Looking for Similarity between Portuguese Ontological Structures

Marcirio Silveira Chaves and Vera Lúcia Strube de Lima

Pontifícia Universidade Católica do Rio Grande do Sul
Faculdade de Informática
Programa de Pós-Graduação em Ciência da Computação - PPGCC
Av. Ipiranga, 6681 - Partenon - Porto Alegre - RS
CEP 90619-900
{mchaves, vera}@inf.pucrs.br

Abstract. The automatic mapping between Ontological Structures (OSs) has been a continuous concern as a task of integration and reuse of knowledge. In order to accomplish part of this task, similarity measures have been used. This paper describes an ongoing work that make use of the String Matching (SM) similarity measure proposed in [1], applied to Portuguese OSs, aiming to finding lexically similar terms. In addition, we present preliminary results for the Lexical Similarity (LS) measure proposed here. These results concern the validation phase of our measure and they are analyzed in detail through three case studies.

1 Introduction

Nowadays, studies that focus the mapping between Ontological Structures (OSs) still keep a considerable amount of manual work. However, the manual execution of this mapping is quite tedious and slow, so it is important to automate, at least partially, the mapping process. Recent proposals [2,3] that address this mapping are described as semi-automatic, since they still lack techniques allowing the full automation of this process.

In this work, OS is understood as a set of pre-defined terms explicitly connected by semantic relations, in a format readable by humans and machines. This notion includes collections of vocabularies as well as concepts.

The task of mapping one OS_A into another OS_B plays a central role on the reuse of available OSs. Ding and Foo [4] mention that the mapping is concerned with the expansion and combination of existing OSs. For example, in the context of information retrieval, as similar terms are found between OSs, a system can browse through combined OSs. This kind of approach [5] helps improving the recall and precision of user queries results.

For Prasad, Peng and Finin [6], mapping OS_A into OS_B consists of a process where, for each concept in OS_A a correspondent concept or term with similar semantics has to be found, in OS_B. If there is no correspondence for such concept or term in OS_B, it is not mapped. In order to help users or systems find similarities between OSs, similarity measures are used.

The semantic similarity measures in [7,8], for example, are based on the content of information of each term, according to a corpus. This content is determined as the number of occurrences of a term, or any child term, in the same hierarchy in a corpus. In our work we do not use a corpus but we apply similarity measures to terms belonging to hierarchies of OSs. However, in a first moment, we work with lexical similarity without concerning about the position of the term in the hierarchy.

Several efforts have been reported in the literature to mapping OSs in English language [2,3,6,9] and in German language [1]. However, works that deal with Portuguese OSs have not been found.

We applied the String Matching measure [1] to Portuguese OSs and several inconsistent results could be observed. So, we worked out the proposal of a measure that uses a stemming algorithm to get better mappings between Portuguese terms, including multiword terms.

This paper is further organized as follows. Section 2 describes works that are related to ours. The application of the String Matching measure to Portuguese OSs is described in Section 3. Section 4 details the similarity measure proposed in this paper and it examines the validation phase for this measure. A discussion on the results of this validation phase is described in Section 5. Finally, in Section 6 we give an outlook on future work.

2 String Matching

Maedche and Staab [1] present a two layer approach, first lexical and then conceptual, to measure the similarity between terms of different OSs. At the lexical level, Maedche and Staab consider the Edit Distance (ED) formulated by Levenshtein [10]. This distance contemplates the minimum number of insertions, deletions or substitutions (reversals) necessary to transform one string into another using a dynamic programming algorithm. For example, ED(`computador`, `computadores`) is 2, because two operations of insertion transform the original string `computador` into `computadores`. The contribution of Maedche and Staab consists of the String Matching (SM) measure given by:

$$SM(T_i, T_j) := max\left(0, \frac{min(\mid T_i \mid, \mid T_j \mid) \; - \; ED(T_i, T_j)}{min(\mid T_i \mid, \mid T_j \mid)}\right) \; \in \; [0, 1] \; . \qquad (1)$$

$SM(T_i, T_j)$ measure calculates the similarity between two terms (Ti, Tj). The length of the shortest term is represented by $min(|Ti|, |Tj|)$. For example, to obtain the similarity between the terms (`computador`, `computadores`) the minimum length is 10 and the value of $ED(Ti, Tj)$ is 2. Thus, the resulting value for SM(`computador`, `computadores`) is 0.8.

The shortest length is considered in the numerator as well as in the denominator of equation 1, what allows pondering the number of changes appearing in the shortest term. The SM measure always returns a value between 0 and 1,

where one stands for perfect match and zero indicates absence of match. Maedche and Staab used German language OSs, specifically in the tourism domain, for their experiments.

3 Applying of SM Measure to Portuguese

We applied the SM measure to Portuguese OSs which come from two distinct sources, the first one from Brazilian Senate (OS_A) and the second one from the São Paulo University - USP (OS_B). The terms appearing in these OSs can belong to one of two groups: single-word terms and multiword terms.

When calculating the similarity by using the SM measure it is important to establish a threshold to detect similar terms. In our experiments were adopt the value 0.75 as a threshold, that is, terms that present similarity values equal or above 0.75 are considered similar, otherwise they are not. This threshold has been used in other works already [1,11].

In Table 1 we present an extract of the results of application of SM measure to Portuguese OSs.

Table 1. Extract of mappings obtained with the application of SM measure to Portuguese OSs

Line	OS_A	OS_B	SM
1	nepotismo	erotismo	0.75
2	realidade	dualidade	0,78
3	criacaoDeEquino	criacaoDeSuinos	0.80
4	rendaPermanente	dentePermanente	0.80
5	datasEspeciais	mapasEspeciais	0.86
6	caminhao	caminhoes	0.62
7	profissao	profissoes	0.67
8	perversaoSexual	perversoesSexuais	0.67
9	embarcacao	embarcacoes	0.70
10	comunicacaoDigital	comunicacoesDigitais	0.72

The pairs of terms in lines numbered from 1 to 5 in Table 1 have distinct semantics, but they are considered similar by SM measure, while the pairs in lines 6 to 10 have the same meaning if number information is not considered distinctive, but they are considered different by SM measure. These inconsistent results occur for single-word terms and also for multiword terms.

Besides, many other results generated by SM measure lead to inconsistent mappings, so that we were motivated to propose another measure, seeking for better results comparisons in Portuguese language. This proposal is explained in section 4.

4 Lexical Similarity Measure

An alternative to SM measure could be based on the radicals[1] of the words. Generally, these radicals are the most representative part of a word in Portuguese language, and they can be extracted by a stemmer. We used a stemmer that was specifically developed for Portuguese language by Orengo and Huyck and presented good results when compared [12] to Porter algorithm or when compared to another algorithm developed for Portuguese language in [13].

The measure proposed in this work is called Lexical Similarity (LS) and it is expressed in equation 2.

$$LS(T_i, T_j) = min\{\Delta_{ij}^1, \Delta_{ij}^2, \ \ldots, \Delta_{ij}^k\} \ \in \ [0, 1] \ . \tag{2}$$

In equation 2 the terms are represented by (T_i, T_j), where index i points to the terms in OS_A while index j refers to terms in OS_B. These terms can be constituted by one word only or by more than one word as well. LS measure, in opposition to SM measure, considers only the radical of each word, and not the complete word with all its characters. The symbols Δ represent the value obtained by SM measure according to the following conditions:

$$\Delta_{ij}^k = \begin{cases} SM(Rad_i^k, Rad_j^k) & if \ ED = 0 \\ SM(Rad_i^k, Rad_j^k) - 0.1 & if \ ED = 1 \\ SM(Rad_i^k, Rad_j^k) - 0.2 & if \ ED = 2 \\ 0 & if \ ED \geq 3 \end{cases} \tag{3}$$

The radical of a word making part of a term (T) is represented by Rad_i^k, where the index k indicates the position of a word in the term and i indicates the OS which the term belong to. When the terms (T_i, T_j) are made of different words, the index k goes on until the amount of words of the term with the minimum number of words. LS measure calculates the similarity between each pair of radicals Rad_i^k, Rad_j^k contained in the terms being mapped.

The final result returned by LS measure is the minimum value generated by equation 3. This value depends on the Edit Distance (ED), which counts the number of insertions, changes or deletions should occur to transform, in the case of LS, a `radical` into another. The final result generated by LS is the minimum value returned by SM measure according to the conditions shown in equation 3.

The radical of a term owns a strong semantic weight in Portuguese. So, the result obtained by ED is decremented according to conditions stated in equation 3. The highest the result generated by SM is, the highest is the penalty used. The penalty values (0.1 and 0.2) were detected from preliminary studies with of SM measure. We assume that, if ED ≥ 3 the value returned by SM is zero. This decision reflects that three or more changes in the radical of a word correspond to a low degree of similarity between the terms being mapping.

[1] The term radical used in this article represents the initial character string of a word and not necessarily the linguistic concept of radical.

For example, in order to verify the similarity between the terms `amazoniaOriental` e `amazoniaOcidental`, the words of each term are processed by a stemming algorithm, producing the following statement:

$$LS(amazoniaOriental, amazoniaOcidental) = min\{SM(amazon, amazon), SM(orient, ocident)\} .$$

$SM(amazon, amazon)$ is 1, while $SM(orient, ocident)$ can be solved as:

$$max\left(0, \frac{6 - 2}{6}\right) = 0.67 .$$

Since in this case $ED = 2$, the penalty to be applied to the value obtained is 0.2. So, the resultant similarity value is 0.47:

$$LS(amazoniaOriental, amazoniaOcidental) = 0.47 .$$

We did not find any other work in the literature that presents a study on semantic weighting each single-word in a multiword term. In our proposal, as the reader can observe, words with the least lexical similarity value may perform an important role on similarity detection between the terms.

After presenting LS measure, we show in the next section the experiments accomplished for the validation of this measure.

4.1 Validation of the Lexical Similarity Measure

The experiments carried out with Portuguese OSs include a validation of LS measure, followed by its evaluation. In order to accomplish these experiments, terms of the OSs were split into two groups. The first one includes single-word terms, and the second one includes multiword terms. The terms of OS_A were separated into two groups for each phase, while OS_B keeps all its terms during both validation and evaluation phases. The terms were placed in alphabetic order and an algorithm was created to randomly distribute them along validation and evaluation experiment groups.

Table 2 shows the amount of terms for each OS according to this distribution, and it shows the global amount of terms used in our experiments. In this paper we focus on the experiments accomplished during the validation phase.

In validation phase, we used 1,824 single-word terms of the Senate OS, while the USP OS remained with the original 7,039 single-word terms. We used 4,701 multiword terms of Senate OS and kept 16,986 multiword terms of USP.

The aim of the experiments during the validation phase was to observe and to analise the behavior of LS measure when applying it to Portuguese terms, as an input to the refinement of this proposal in order to tune our method to the next (evaluation) phase. To carry out these experiments we used the threshold 0.75 and we split the terms as shown[2] in Table 3.

[2] Pairs of terms with similarity value 1 are not considered in these experiments. Terms considered different by both SM and LS measures are also discarded.

Table 2. Distribution of terms of Portuguese OSs

OS	Term Category	Amount of Terms		Total Term Category	%	Total of Terms
		Validation	Evaluation			
Senate	single-word	1,824	1,823	3,647	28	13,049
	multiword	4,701	4,701	9,402	72	
USP	single-word	-	-	7,039	29	24,025
	multiword	-	-	16,986	71	

Table 3. Case studies in validation phase

Case #	Conditions	
1	SM ≥ 0.75	LS ≥ 0.75
2	SM ≥ 0.75	LS < 0.75
3	SM < 0.75	LS ≥ 0.75

Table 3 presents the cases of combinations between SM and LS similarity measures. These cases are explained as follows:

1. Case #1 refers to "agreement" of similarity calculation between SM and LS measures, it concerns the terms where both measures detect similarity;
2. Case #2 refers to the terms that are considered similar by SM measure but not similar by LS measure;
3. Case #3 groups the terms that are not considered similar by SM measure but are considered similar by LS measure.

Table 4 depicts the results generated on validation phase, to each of the cases in Table 3.

Table 4. Amount of mappings obtained for each case depicted in Table 3

Case #	Single-word terms	Multiword terms	Total
1	53	18	71
2	1,026	1,608	2,634
3	45	10	55
Total	1,124	1,636	2,760

According to Table 4, 2,760 pairs of terms were considered similar by one of measures (SM measure or LS measure). The reader may note that most of the pairs of terms (2,634 pairs, or, more than 95%) analyzed during the validation phase, are in the second case so, SM measure considers them similar and LS measure considers them different.

5 Discussion on the Results

In this section we analyze each case depicted in Table 3, aiming to refining the LS measure before going through evaluation phase.

5.1 Case #1: Agreement between SM and LS Measures

In this case, both measures consider the terms being compared as similar. An extract of those pairs of terms is presented in Table 5.

For the experiments with multiword terms, OSs were first preprocessed in order to eliminate blanks. Besides, the first character of each word was capitalized, except for the first word of the term. This procedure is necessary to allow us detecting the beginning of each word. This preprocessing has also been used for other experiments in [3,1].

Table 5. Extract of pairs of terms considered similar by both SM and LS measures

OS_A	OS_B	SM	LS
cartilha	**p**artilha	0.88	0.76
dolarizacao	**p**olarizacao	0.91	0.80
emigracao	**i**migracao	0.89	0.77
fitologia	**m**itologia	0.89	0.76
ginecologia	**s**inecologia	0.91	0.79
matrimonio	**p**atrimonio	0.90	0.79
ovinocultura	**b**ovinocultura	0.92	0.79
acumulacaoDeAcoes	cumulacaoDeAcoes	0.94	0.77
mercado**M**obiliario	mercado**I**mobiliario	0.88	0.76

Although they were considered similar, each pair of terms in Table 5 seems to lack semantic similarity. Or, these pairs of terms present a high lexical similarity, but their meanings are quite different.

As shown in Table 5, for most of the pairs of terms, just the first character of the string is different. In fact, for Portuguese language, the semantic weight of the first characters in a term is strong, which gives rise to the "first letter heuristic" proposed to deal with this kind of situation. This heuristic is stated as follows:

$$If\ Rad[1]_i^k \neq Rad[1]_j^k\ then\ SM(Rad_i^k, Rad_j^k) = 0$$

According to LS measure presented in equation 2, let the index inside the brackets be the position of the first character in the radical of a word in a term. If the two radicals Rad_i^k, Rad_j^k being compared have their first letter different, the value returned by SM measure is zero. Consequently, LS is zero, too.

Table 5 presents some results for case #1 without application of this heuristic. When applying the first letter heuristic to these terms, similarity values are zero and, consequently, no mapping is created. In this table two multiword terms are presented. In this case, the first letter heuristic is applied to all radicals making part of the multiword term. For `mercadoMobiliario` and `mercadoImobiliario`, the heuristic was applied to the second radical, once there was no different character for the first one.

Concerning multiword terms, 17 pairs of terms were detected as similar in case #1. Through the extract presented in Table 6 it is possible to identify number variation[3] between the terms. This difference disappears with the use of a stemmer, since each word of the term is reduced to its radical.

Table 6. Multiword terms with number variation considered similar by SM and LS measures

OS_A	OS_B	SM	LS
acumulacaoDeAcoes	cumulacaoDeAcoes	0.94	0.77
bicho-da-seda	bichos-da-seda	0.92	0.82
competicaoEsportiva	competicoesEsportivas	0.79	0.79
condicoesEconomicas	condicaoEconomica	0.76	0.76
condicoesSanitarias	condicaoSanitaria	0.76	0.76
construcaoMetalica	construcoesMetalicas	0.78	0.79
criacaoDeCaracol	criacaoDeCaracois	0.88	0.76
descobertaEExploracao	descobertasEExploracoes	0.81	0.79
expedicaoCientifica	expedicoesCientificas	0.79	0.77
exposicaoInternacional	exposicoesInternacionais	0.77	0.77
instituicaoFinanceira	instituicoesFinanceiras	0.81	0.80
instituicaoPolitica	instituicoesPoliticas	0.79	0.80
religiaoPrimitiva	religioesPrimitivas	0.76	0.76

Table 6 shows 13 (76,5% of this group) selected pairs from the 17 multiword pairs of terms considered similar by SM and LS measures. This is a circumstantial evidence that SM and LS can treat multiword terms with number variation in a consistent way.

Terms with number variation are frequently found in OSs because, as observed by Noy and McGuinness [14], people model knowledge using a standard: all terms in singular or all terms in plural. The problem occurs when a knowledge engineer has to map an OS modelled in singular into an OS modelled in plural.

5.2 Case #2

This case presents the pairs of terms considered similar by SM measure and unlike by LS measure. An extract of those pairs is shown in Table 7.

[3] Terms in singular or plural form.

Table 7. Extract of terms case #2

OS_A	OS_B	SM	LS	OS_A	OS_B
mortalidade	moralidade	0.90	0.70	mortal	moral
impunidade	imunidade	0.89	0.65	impun	imun
teologia	geologia	0.88	0.73	teolog	geolog
modelos	modulos	0.86	0.70	model	modul
moveis	moteis	0.83	0.70	movel	motel
areaEstrategica	armaEstrategica	0.93	0.57	areestrateg	armestrateg
cartaDeCredito	cartaoDeCredito	0.93	0.65	cartdecredit	cartadecredit
capacidadeJuridica	incapacidadeJuridica	0.89	0.40	capacjurid	incapacjurid
linguasIndigenas	linguasIndianas	0.87	0	linguindigen	linguindi
mitoPopular	votoPopular	0.82	0.13	mitpopul	votpopul

Table 7 includes single-word and multiword terms. In this extract it is possible to note that all pairs seem dissimilar, even if they were considered similar by SM measure and unlike by LS measure.

In this case #2 the LS measure presented its best performance because it does not consider similar terms with different meanings. We may observe that all the pairs present the same final string, that is, the suffix of the terms is the same. As the LS measure eliminates the suffix and introduces a penalty to modifications on the radicals of the terms, the similarity value is reduced.

On the other hand, terms which present different suffixes are not considered similar by LS. An extract with some of these terms is shown in Table 8.

Table 8. Extract of terms with different suffixes in case #2

OS_A	OS_B	SM	LS	OS_A	OS_B
pintor	pintos	0.83	0.57	pin	pint
corretora	corredor	0.75	0.65	corre	corr
remicaoDeBens	remicaoDaPena	0.77	0.13	remicadebem	remicadapen
livroComercial	livreComercio	0.77	0.73	livrcomerc	livrcomerci
arquiteturaDeRede	arquiteturaDeTerra	0.76	0	arquitetdered	arquitetdeterr
assistenciaMilitar	assistenciaMedica	0.76	0	assistencmilit	assistencmedic

The extracts presented in both Tables 7 and 8 are a circumstantial evidence that SM measure is not adequate to deal with terms in Portuguese, because most of the pairs of terms (about 95%) mapped during validation phase belong to this second case. In addition, the remaining pairs of terms, not presented in Tables 7 and 8 (due to space limitations), can also be considered as inconsistent mappings, according to SM measure.

In Table 8 it is interesting to mention that the pair of terms `livroComercial` and `livreComercio` is not detected as similar by LS because of a stemming failure: the stemming algorithm did not completely remove the suffix of the

word `Comercio`. This situation demonstrates some of the difficulties that appear when detecting similarity between terms in Portuguese. Sometimes, the results returned by the stemmer are not the exact linguistic radical of a word, but the initial string of a word. Terms with this same initial string, like `livro` and `livre`, may present the same stem even assuming distinct meanings.

5.3 Case #3

In this case we present terms that are unlike according to SM measure and are considered similar by LS measure. Initially, we could emphasize the single-word terms with number variation that present inconsistent results generated by SM measure, like those depicted in Table 9.

Table 9. Extract of single-word terms with number variation, considered similar by LS in case #3

OS_A	OS_B	SM	LS
adivinhacao	adivinhacoes	0.73	0.80
caminhao	caminhoes	0.62	0.76
corporacao	corporacoes	0.70	0.79
embarcacao	embarcacoes	0.70	0.79
habitacao	habitacoes	0.67	0.77
profissao	profissoes	0.67	0.77
religiao	religioes	0.62	0.76

Unlike the multiword terms depicted in Table 6, these terms are not considered similar by SM measure while using the threshold 0.75, although the LS measure considers them similar. It is also important to pay attention to the stemming mistakes that occur during this process. Some mistakes concern the change of character `ç` by `c` and the elimination of the $\sim$ character. The OSs worked here, due to an original standardization strategy, did not contain those characters.

Likewise, the stemming mistakes also prejudice the performance of LS measure. We show in Table 10 some terms that were considered similar by LS measure because the stemming algorithm did not completely remove their suffix. Such mistakes are known as "understemming".

In the analysis of Table 10, it is possible to note that the pairs of terms are not similar. When these pairs are manually checked (see last two columns), LS measure presents values below the established threshold (0.75), except for the pair `profissao` and `profissoes`, whose similarity value is 1. In the context of OSs, these terms are effectively similar.

Finally, in the analysis of case #3, we found the multiword terms `auto-estrada` and `auto-estima` detected as similar (0.77) by LS and considered unlike (0.73) by SM. In this situation, LS measure not mapped the terms correctly.

Table 10. Terms that present mistakes produced by the stemming process

OS_A	OS_B	SM	LS	checked LS	checked stem	checked stem
empresario	emprestimo	0.70	0.76	0.73	empres	emprest
inflamaveis	inflacao	0.38	0.76	0.73	inflam	inflac
magistrado	magisterio	0.70	0.76	0.73	magistr	magist
metanol	metabolismo	0.29	0.76	0	metan	metabol
profissao	profissoes	0.67	0.77	1	profiss	profiss
responsabilidade	responsorio	0.27	0.76	0.51	respons	responsor

6 Final Remarks and Future Work

In this paper we present an ongoing research that investigates alternatives to detect similarity between terms in Portuguese OSs. We apply the SM measure to these OSs and present some preliminary results. It was possible to confirm that SM measure alone is not enough to detect consistent similarities. In addition, we present the LS measure and the experiments accomplished during its validation phase. An analysis of each case was carried out, in which we discussed the characteristics of terms where the LS measure presented good performance as well as the cases where this measure did not show to be adequate.

This work is a first effort towards the detection of similar terms between Portuguese OSs. We believe that the problem of semantic similarity is not yet completely treated. However, the LS measure can be used as an initial step in a task of integration. This integration allows the reuse of information, which reflects a concern of the researches toward the semantic web approach.

We are aware that it is necessary to undertake a deeper evaluation of our proposal. This deeper evaluation will be the next step of this work and the results will be compared with those obtained from human similarity detection. Another future work is the application of LS measure to other languages, such as English or Spanish. In these situations a proper stemming algorithm, suitable for each different language, should be used.

Acknowledgements

We thank Viviane Orengo that kindly provided us the stemmer used here. Marcirio Silveira Chaves is supported by the research center HP-CPAD (Centro de Processamento de Alto Desempenho HP Brasil-PUCRS).

References

1. Alexander Maedche and Steffen Staab. Measuring Similarity between Ontologies. In *Proceedings of the European Conference on Knowledge Acquisition and Management - (EKAW-2002). Madrid, Spain, October 1-4*, pages 251–263, 2002.

2. AnHai Doan, Jayant Madhavan, Pedro Domingos, and Alon Halevy. Learning to Map between Ontologies on the Semantic Web. In *Proceedings of the World-Wide Web Conference (WWW-2002), Honolulu, Hawaii, USA*, May 2002.
3. Natalya Fridman Noy and Mark A. Musen. Anchor-PROMPT: Using Non-Local Context for Semantic Matching. In *Proceedings of the Workshop on Ontologies and Information Sharing at the Seventeenth International Joint Conference on Artificial Intelligence (IJCAI-2001), Seattle, WA*, August 2001.
4. Ying Ding and Schubert Foo. Ontology Research and Development Part 2 - A Review of Ontology Mapping and Evolving. *Journal of Information Science*, 28(5):375–388, 2002.
5. Luiz Augusto Sangoi Pizzato and Vera Lúcia Strube de Lima. Evaluation of a thesaurus-based query expansion technique. In *Proceedings of the $6^t h$ Workshop on Computacional Processing of the Portuguese Language - Written and Spoken (PROPOR), Lecture Notes in Computer Science*, Universidade do Algarve-FCHS, Faro, Portugal, June 26-27 2003. Springer-Verlag.
6. Sushama Prasad, Yun Peng, and Timothy Finin. Using Explicit Information To Map Between Two Ontologies. In *Proceedings of the 1st International Joint Conference on Autonomous Agents and Multi-Agent Systems - Workshop on Ontologies in Agent Systems (OAS) - Bologna, Italy. 15-19 July*, 2002.
7. Philip Resnik. Using Information Content to Evaluate Semantic Similarity in a Taxonomy. In *Proceedings of the XI International Joint Conferences on Artificial Intelligence (IJCAI)*, pages 448–453, 1995.
8. Jay J. Jiang and David W. Conrath. Semantic Similarity based on Corpus Statistics and Lexical Taxonomy. In *Proceedings of the International Conference on Research in Computational Linguistics (ROCLING)*. Taiwan, 1997.
9. Natalya Fridman Noy and Mark A. Musen. SMART: Automated Support for Ontology Merging and Alignment. In *Twelfth Banff Workshop on Knowledge Acquisition, Modeling, and Management - Banff, Alberta, Canada*, 1999.
10. Vladimir Levenshtein. Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Cybernetics and Control Theory*, 10(8):707–710, 1966.
11. Marcirio Silveira Chaves and Vera Lúcia Strube de Lima. Similaridade entre Estruturas Ontológicas. In *Proceedings of XVI Brazilian Symposium on Computer Graphics and Image Processing - (SIBGRAPI). I Workshop em Tecnologia da Informação e Linguagem Humana, São Carlos-SP, Brasil (CD-ROM), 12 de Outubro de 2003.*
12. Viviane Moreira Orengo and Christian Huyck. A Stemming Algorithm for Portuguese Language. In *Proceedings of Eigth Symposium on String Processing and Information Retrieval (SPIRE-2001)*, pages 186–193, 2001.
13. Marcirio Silveira Chaves. Um Estudo e Apreciação sobre Dois Algoritmos de Stemming para a Língua Portuguesa. IX Jornadas Iberoamericanas de Informática. Cartagena de Indias - Colômbia (CD-ROM), 11-15 de Agosto 2003.
14. Natalya Fridman Noy and Deborah L. McGuinness. Ontology Development 101: A Guide to Creating Your First Ontology. Technical report, Stanford Knowledge Systems Laboratory, Technical Report KSL-01-05 and Stanford Medical Informatics, Technical Report SMI-2001-0880, 2001.